615 Final

Ziyang Lin

2022-12-14

This dataset describes the travel times between origin and destination pairs on a single line. It contains data up to the most recent completed quarter for 2022. These travel times are calculated from the departure time at the origin stop to the arrival time at the destination stop. We randomly picked one week per month, excluding the first and last seven days in the month.

```
## # A tibble: 6 x 12
##   service_~1 from_~2 to_st~3 route~4 direc~5 start~6 end_t~7 trave~8 month  week
##   <date>       <dbl>   <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 2021-10-08   70111   70107 Green-B       0   30335   30489     154    10    41
## 2 2021-10-08   70111   70107 Green-B       0   29583   29775     192    10    41
## 3 2021-10-08   70111   70107 Green-B       0   36628   36802     174    10    41
## 4 2021-10-08   70111   70107 Green-B       0   37307   37411     104    10    41
## 5 2021-10-08   70111   70107 Green-B       0   37433   37643     210    10    41
## 6 2021-10-08   70111   70107 Green-B       0   36355   36571     216    10    41
## # ... with 2 more variables: weekdays <chr>, trip <fct>, and abbreviated
## #   variable names 1: service_date, 2: from_stop_id, 3: to_stop_id,
## #   4: route_id, 5: direction_id, 6: start_time_sec, 7: end_time_sec,
## #   8: travel_time_sec
```

```
##   service_date         from_stop_id      to_stop_id       route_id
##  Min.   :2021-10-08   Min.   : 70110   Min.   : 70107   Length:15550482
##  1st Qu.:2022-01-22   1st Qu.: 70153   1st Qu.: 70153   Class :character
##  Median :2022-04-16   Median : 70179   Median : 70179   Mode  :character
##  Mean   :2022-04-04   Mean   : 72859   Mean   : 72778
##  3rd Qu.:2022-06-17   3rd Qu.: 70226   3rd Qu.: 70227
##  Max.   :2022-09-23   Max.   :170141   Max.   :170141
##
##   direction_id    start_time_sec   end_time_sec   travel_time_sec
##  Min.   :0.0000   Min.   :14402   Min.   :17116   Min.   :    0.0
##  1st Qu.:0.0000   1st Qu.:36982   1st Qu.:37800   1st Qu.:  288.0
##  Median :0.0000   Median :53728   Median :54562   Median :  665.0
##  Mean   :0.4726   Mean   :53893   Mean   :54697   Mean   :  804.5
##  3rd Qu.:1.0000   3rd Qu.:70325   3rd Qu.:71133   3rd Qu.: 1198.0
##  Max.   :1.0000   Max.   :93588   Max.   :93596   Max.   :13043.0
##
##      month             week         weekdays
##  Min.   : 1.000   Min.   : 4.00   Length:15550482
##  1st Qu.: 4.000   1st Qu.:16.00   Class :character
##  Median : 6.000   Median :24.00   Mode  :character
##  Mean   : 6.447   Mean   :26.59
##  3rd Qu.: 9.000   3rd Qu.:38.00
```

```
##  Max.   :12.000   Max.   :52.00
##
##                   trip
##  From 70159 to 70157:   41675
##  From 70157 to 70155:   40659
##  From 70156 to 70158:   40656
##  From 70159 to 70155:   40512
##  From 70154 to 70156:   40087
##  From 70154 to 70158:   39416
##  (Other)            :15307477
```

```
## [1] 15550482       12
```

We first take a look at the first six rows of the dataset. Then, the quickly summarize each variable of the dataset. After that the dimension of the dataset in terms of number of rows and number of columns is showed. Below is the data dictionary:

service_date: Date for which travel times should be returned.

route_id: GTFS-compatible route for which travel times should be returned.

direction_id: GTFS-compatible direction for which travel times should be returned. from_stop_id: GTFS-compatible stop representing the origin stop in a pair.

to_stop_id: GTFS-compatible stop representing the destination stop in a pair. start_time_sec: The time associated with the departure event of the vehicle from the origin stop of the pair.

end_time_sec: The time associated with the arrival event of the vehicle to the destination stop of the pair.

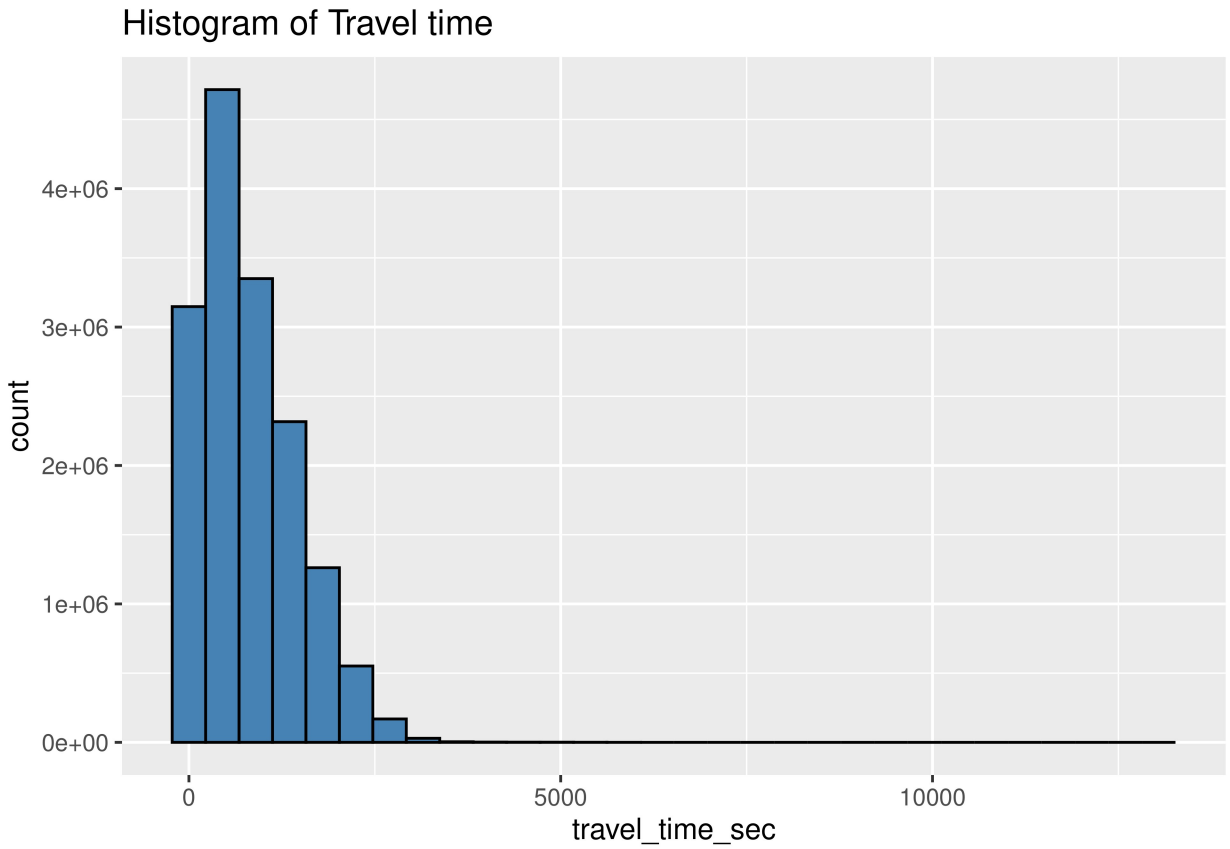
travel_time_sec: Difference between start_time_sec and end_time_sec. The actual travel time between the origin stop and the destination stop, in seconds.

week: random weeks we pick

month: month in a year

weekdays: exactly which weekdays is that day

trip: departure and arrival stop for the trip



A histogram is created to show the distribution. From the plot, we could find out that most of the travel times between stops are less than 2,500 seconds.

```
##    service_date    from_stop_id      to_stop_id        route_id    direction_id
##               0               0               0               0               0
##  start_time_sec    end_time_sec travel_time_sec           month            week
##               0               0               0               0               0
##        weekdays            trip
##               0               0
```

The total number of missing values in each column of the dataset is counted. We could find out that there are zero missing values in each column.

For EDA and shiny, I did the T part. The T part has most completed data for both date and variables. For the rest of services, the data are relatively not completed and there are some variables I don't understand. For this project, I learned things like how to deal with large datasets, apply some of the variables to the maps, using leaflet and google api.